DOCUMENT RESUME

ED 129 849 TM 005 469

AUTHOR Morgan, Penelope; And Others
TITLE It's the Metric That Counts; or, Criterion Referenced Schizophrenia.
PUB DATE [Apr 76]
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Classroom Materials; *Criterion Referenced Tests; Elementary Secondary Education; *Evaluation Criteria; Evaluation Methods; *Program Evaluation; Test Construction; *Test Reviews; Test Selection

ABSTRACT

Fourteen criteria used to evaluate criterion referenced tests were assigned two sets of weights reflecting the characteristics of tests designed as a classroom resource or for program evaluation. Twenty eight currently available criterion referenced tests were rated against the criteria using each set of weights. A comparison of the scores obtained with each weighting system yielded significant differences. The findings of this study support the view that the same criteria cannot be used for all purposes and that therefore, a criterion referenced test must be developed, validated, and evaluated in terms of the purpose for which it is intended. (Author/BW)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED129849

It's the Metric That Counts

or

Criterion-Referenced Schizophrenia

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Penelope Morgan, Jacqueline Kosecoff
Clint Walker, J. Ward Keesling

Paper presented at the annual meeting of the
American Educational Research Association

San Francisco, April 1976

TM005 469

It's the Metric That Counts
or
Criterion-Referenced Schizophrenia

The academic and lay communities have, for over two decades, been trying to develop educational measurement systems which accurately gauge and monitor the performance by an individual or group of individuals on a given set of instructional tasks, or behavioral objectives, in such a way as to yield pertinent and useable information to those persons providing the educational experience. The results of these efforts are called criterion-referenced tests (CRTs), and they are deliberately constructed to permit score interpretations in terms of pre-specified performance standards. That is, CRTs can be used to describe what an individual or group of individuals can or cannot do (Glaser and Nitko, 1971; Popham and Husek, 1969).

The reaction to CRTs has been enthusiastic from the start. Because they provide score interpretations in terms of the achievement of specific and measureable skills and behaviors, CRTs have had appeal to those directly responsible for the education of students and the development and evaluation of educational programs. They also have had appeal to teachers who found the results of standardized tests inadequate to assist them in planning lessons, and to many educators and psychologists who judged standardized, norm-referenced tests to be unfair and even biased against individuals from under-privileged and minority groups. Finally, because the criterion-referenced approach was new, people have seen it as an opportunity to improve on some of the mistakes they perceived to be built into, or confused with, norm-referenced testing (Kosecoff and Fink, 1976b).

CRT's popularity and sanction by theoreticians and practitioners has led to their frequent use for instructional diagnosis and placement and for measuring student achievement on educational tasks or objectives. In addition, CRTs are being suggested or used for other purposes like the evaluation of educational programs and the National Assessment of Educational Progress (Wilson, 1974). In fact, many recently-issued requests for proposals from state and federal agencies to evaluate educational programs have specifically required prospective contractors to justify their selection of standardized rather than CRT measures.

The question then arises if the same CRTs can be used for classroom and evaluation purposes, and if not, what are the characteristics of CRTs best suited for each situation? Until recently, the answer to this question would probably have been yes, that the same CRTs can be effectively used for a variety of purposes. One reason for this answer is that publishers of CRTs have simultaneously recommended their tests for use in diagnosing learning problems, placing students in a curriculum, measuring progress, and evaluating instructional programs (Kosecoff, Klein, and Morgan, 1976). A second reason is that in the CRT literature, most educators and psychometricians have not distinguished between CRTs intended for different purposes either in terms of differences in physical characteristics or psychometric properties (Alkin et al, 1974). Third, by applying the same criteria and the same weights to all CRTs, efforts to evaluate CRTs also have not recognized different types of CRTs (Nafziger et al., 1975).

Recently, some test developers, theoreticians and users have begun to study the characteristics of CRTs designed for different purposes. For example, the Instructional Objectives Exchange is developing, for school

districts, customized tests that differ considerably in format and development from other commercially available CRTs (Popham, 1976), and in a recent study for System Development Corporation (SDC), Kosecoff and Fink (1976a) concluded that no commercially available CRT was appropriate for use in SDC's longitudinal, nation-wide study of compensatory education.

The purpose of this paper is to lend support to the view that different decision purposes require CRTs with different physical and theoretical properties, and that therefore, CRTs must be evaluated in terms of the context in which they will be used. To do this, two situations in which CRTs are being used, the classroom resource and the program evaluation context, were considered and a sample of criteria that have been used or are being used to evaluate CRTs were assigned weights appropriate to each context. Then, twenty eight commercially published CRTs were rated using each set of weights.

The Classroom Resource Context

One of the original and most prevalent uses of CRTs is as an instructional aide to teachers in planning, evaluating, and individualizing instruction. CRT results can be used by teachers to identify individual's or group's learning problems, to place students in a curriculum, to organize instructional groups, to monitor student's progress throughout a curriculum, and to measure the extent to which students mastered specific instructional objectives. In this context, CRTs can be seen primarily as tools used by teachers to obtain information for classroom management (Anatasi, 1968).

The Program Evaluation Context

CRTs can be used to provide information useful in evaluating educational

programs. This evaluation information can be used to modify and improve a still-developing program or to ascertain the effectiveness and worth of a mature program (Kosecoff and Fink, 1976b). In this context, CRTs can be seen primarily as tools used by evaluators to provide program developers and/or program sponsors with information about the nature and quality of a program's goals, outcomes, impacts, and costs.

Evaluation Criteria

In order to investigate differing characteristics of CRTs intended for more than one purpose, twenty eight currently available CRTs were reviewed using a dually-weighted set of fourteen criteria. The criteria chosen for the review were selected, based on a literature review, from criteria lists that are currently being used or have been used to evaluate educational tests (Hoepfner et al., 1970; Klein and Kosecoff, 1973; Kosecoff, Klein and Morgan, 1976; Kosecoff and Fink, 1976b; Buros, 1972). For this review, two sets of weights were assigned to each criteria, the first set of weights reflecting characteristics desirable of CRTs intended for use as a classroom resource and the second set of weights reflecting characteristics desirable of CRTs intended for use in program evaluation.

An explanation of each criterion, its rating categories, and the assigned weights are:

(1) Overlap of objectives across grade levels. Some or all of the test's objectives must be measured at each grade level. This criterion is particulary important in a program evaluation context in order to make comparisons across grade levels or over time in terms of common educational objectives or skills.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Overlap of objectives across grade levels	no overlap	0	0
	overlap in goals	1	2
	some overlap in objs.	1	2
	overlap in goals and some objectives	2	3
	total overlap	1	2

(2) Grade level coverage. The number of grade levels for which some form of the CRT is appropriate. This criterion is slightly more important for CRTs intended for use in evaluations, since it is frequently necessary to interpret trends or outcomes in this context at more than one grade level.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Grade level coverage	just one grade	0	0
	several grades (K-3 or 4-6)	1	2
	K-6	2	3

(3) Number of test forms. The number of different tests or subtests (measuring different objectives) intended for a given grade level. Due to constraints related to test administration and the time available for testing in an evaluation context, it is desirable to have a limited number of test forms at each grade level. Just one test per grade level is preferred in order to avoid problems with reliability that can arise when several test forms are combined. On the other hand, in a classroom context, many short test forms, each measuring just a few very specific instructional objectives might be preferred.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Number of test forms per grade level (not alternate forms)	just one test form	0	3
	2-9 tests	2	1
	10 or more tests	3	0

(4) Special equipment needed for test administration. In view of the variation in equipment from school to school and in an effort to reduce costs, test administration in an evaluation context should not require any special equipment (like cassettes or visual aids). In a classroom context it can be useful to make use of this equipment for example, in individualizing testing. However, it is preferred that special equipment be available as an option and not be mandatory.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Special equipment for test administration	none needed	2	3
	yes, but optional	1	1
	yes, mandatory	0	0

(5a) Average time for administering a single test. (Criterion 5a applies just to CRTs with more than one test per grade level, excluding alternate test forms). The amount of time required, on the average, to administer a single test at a given grade level. For this criterion, the principle "the less testing time the better," applies to CRTs intended for use in either context, however, evaluations usually have more stringent limits on the total amount of time for testing.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Average time for administering a single test (if more than one test per level)	10 or less minutes	3	1
	10-20 minutes	2	0
	20 or more minutes	1	0

(5b) Average time for administering all tests at a grade level (excluding parallel forms). The amount of time needed to administer all tests (excluding parallel test forms) at a grade level. For use in an evaluation context, it is desirable that the CRT be designed to be completed within a class period.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Average time for testing for all forms at a grade level (not alternate forms)	45 or less minutes	0	2
	45-90 minutes	0	1
	90 or more minutes	1	0

(6) Machine/self-scoring. This refers to whether the test can be machine scored, hand scored, or both. Sending CRTs to a publisher for machine-scoring, can waste valuable time in a classroom context, and consequently hand-scoreable tests are preferable. However, in an evaluation context where it is not infrequent to test thousands of students, machine scoring is both efficient and accurate.

Metric

B

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Machine/self scoring	just machine	0	3
	just self	3	1
	both	3	3

(7) Score interpretation. CRT scores are reported in terms of the level of performance achieved for each objective measured by a CRT. Some CRTs report scores as the number of items correctly answered per objective or as mastery of an objective, where mastery is an arbitrarily defined level of performance. CRTs using these score interpretation schemes usually are intended for use in a classroom context and have very specific and operationally-defined objectives that can be directly measured by the test items. Other CRTs report scores in terms of true level of performance on an objective, referring to the portion of the total universe of items that could be correctly answered, or in terms of empirically-determined mastery levels. CRTs using this type of score interpretation scheme usually are intended for use in an evaluation context and have more generally-stated objectives for which it is difficult to assume that achievement of the test items necessarily reflects achievement of the larger objective.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Score interpretation	# items per obj. or arbitrary mastery	2	0
	true score or empirically-based mastery	3	3

(8) Curriculum match. Some CRTs are designed for use with a specific educational program (Baker, R.L., 1972; Skager, 1973). CRTs that closely match a curriculum have objectives and test items that are associated with a particular set of educational materials and techniques. CRTs with a smaller match to a curriculum contain objectives and test items that are not necessarily associated with the specific skills or content of an educational program. However, they still may have been developed from several educational programs and consequently, have objectives and items that reflect the bias inherent in these programs. Conversely, CRTs with no match to a curriculum are based on objectives and test items that are independent of any educational program, and, therefore, can be used to compare several different educational programs. In the classroom context it is useful to use a CRT that is closely matched to a curriculum, while in an evaluation context involving comparisons, the opposite it true.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Curriculum match	none	1	3
	somewhat	2	2
	close	3	1

(9) Who can interpret scores. This criterion refers to whether a teacher can interpret the CRT's score or if an expert is needed. Clearly, in the classroom context a teacher should be able to interpret scores, while the evaluator is, by training, skilled in test interpretation.

Metric

10

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Who can interpret scores	teacher	3	1
	specialist	0	0

(10) Formal field test. It is a well excepted fact that any test should be carefully validated. However, a carefully documented field test report is particularly important in a program evaluation context because far-reaching decisions may be made on the basis of the evaluation findings.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Formal field test	none mentioned	0	0
	mention but little documentation	0	0
	yes, restricted scope	1	2
	yes, national scope	2	3

(11) Stability/number of items per objective. A determination of the number of items that must be tested in order to obtain a stable score on an objective. For a CRT used in a classroom context, it is necessary to establish stability for individual examinees, while for a CRT used in an evaluation context, it is desirable to determine stability for groups of examinees as well as individuals. (Note, fewer items are probably needed to obtain stable scores for a group than for an individual.)

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Stability/number of items per objective	stability for indivs.	2	2
	stability for groups	1	2
	both	2	2

(12) Sensitivity to instruction. This criterion refers to how well a CRT distinguishes between those who have and those who have not benefited from instruction. In a classroom context, a teacher is most likely interested in using a CRT to determine if students mastered one or more instructional objectives. It is important therefore, that the CRT be sensitive to the teacher's instruction.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Sensitivity to instruction	study not documented	0	0
	study documented	2	1

(13) Subject area comprehensiveness. Some CRTs attempt to cover a given subject area by measuring many specific objectives. Other CRTs attempt measuring just a limited number of objectives. However, the objectives are selected so that performance on them is generalizable to other skills associated with the subject area. In an evaluation context, it is desirable to use the second kind of CRT in order to ensure a manageable amount of information and to meet testing time constraints. In a classroom context, the first type of CRT is probably more useful since students can be tested on instructional objectives as they are introduced.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Subject area comprehensiveness	test measures more than 20 objectives	2	1
	test samples objs. but sample has predictive validity	1	2

(14) Availability of comparative information. In addition to a CRT interpretation, norm-referenced interpretations can also be provided. Comparative data describing how other students have performed on the same objective are useful in both the classroom resource and program evaluation contexts to enhance the meaningfulness of CRT scores.

CRITERION	RATING CATEGORIES	WEIGHTS CLASSROOM RESOURCE	WEIGHTS PROGRAM EVALUATION
Availability of comparative information	available	2	2
	not available	0	0

A complete rating scale can be found in Figure 1.

see page 13 for Figure 1

Results of the Review

Twenty eight CRTs currently being marketed by major publishers of educational tests were reviewed. The names of the systems reviewed and their respective publishers are listed in Table 1.

see page 14 for Table 1

Copies of the CRTs were obtained from the Center for the Study of Evaluation's test library. Each CRT was independently reviewed twice using the criteria selected for this purpose and discrepancies were resolved by the reviewers. Any remaining questions, that is, those resulting from unclear or insufficient information were followed-up by a phone call to the publisher.

For the review, when sufficient information could not be obtained to rate a criterion, a score of zero was assigned. Review scores were computed

Figure 1: Rating Categories

CRITERIA	RATING CATEGORIES	POINTS CLASSROOM	POINTS EVALUATION
1. Overlap of objectives across grade levels	no overlap	0	0
	overlap in goals	1	2
	some overlap in objectives	1	2
	overlap in goals and some objectives	2	3
	total overlap	1	2
2. Grade level coverage	just 1 grade	0	0
	several grades (K-3 or 4-6)	1	2
	K-6	2	3
3. Number of test forms per grade level (not alternate forms)	just 1 test form	0	3
	2-9 tests	2	1
	10 or more tests	3	0
4. Special equipment for test administration	none needed	2	3
	yes, but optional	1	1
	yes, mandatory	0	0
5a. Average time for administering a single test (if more than one test form per level)	< 10 min.	3	1
	10-20 min	2	0
	>20 min	1	0
	not applicable	0	0
5b. Average time for administering all tests at a grade level (excluding parallel test forms)	<45 min.	0	2
	45-90 min.	0	1
	>90 min	1	0
6. Machine/Self scoring	just machine	0	3
	just self	3	1
	both	3	3
7. Score interpretation	# items per obj. or arbitrary mastery	2	0
	true score or empirically-based mastery	3	3
8. Curriculum match	none	1	3
	somewhat	2	2
	close	3	1
9. Who can interpret scores	teacher	3	1
	specialist	0	0
10. Formal field test	none mentioned	0	0
	mention but little documentation	0	0
	yes, restricted scope	1	2
	yes, national scope	2	3
11. Stability-number of items per objective	stability for individuals	2	2
	stability for groups	1	1
	both	2	2
12. Sensitivity to instruction	study not documented	0	0
	study documented	2	1
13. Subject-area comprehensiveness	test measures many objectives (>20 per level)	2	1
	test samples objectives but sample has predictive validity	1	2
14. Availability of comparative information	available	2	2
	not available	0	0

Table 1

Alphabetical listing of publisher and test systems reviewed
American Guidance Service
Key Math - Diagnostic Arithmatic Test
Woodcock Reading Mastery
American Testing Company
Mathematics Inventory Probe
Reading Inventory Probe I
CTB/McGraw-Hill
Comprehensive Tests of Basic Skills (CTBS/S) - Mathematics
Comprehensive Tests of Basic Skills (CTBs/S) - Reading
Diagnostic Mathematics Inventory
Objectives-Referenced Bank of Items and Tests (ORBIT)
Prescriptive Reading Inventory
Educational and Industrial Testing Service
Tests of Achievement in Basic Skills (TABS) - Mathematics
Tests of Achievement in Basic Skills (TABS) - Reading
Educational Development Corporation
Individualized Criterion-Referenced Testing - Mathematics
Individualized Criterion-Referenced Testing - Reading
Harcourt Brace Jovanovich
Skills Monitoring System - Reading (not yet available)
1973 Stanford Mathematics Tests
1973 Stanford Reading Tests
Houghton - Mifflin
Individual Pupil Monitoring System - Mathematics
Individual Pupil Monitoring System - Reading
Instructional Objectives Exchange
Objectives Based Test Sets - Mathematics
Objectives Based Test Sets - Reading
National Evaluation Systems
Comprehensive Achievement Monitoring (CAM) - Mathematics
Comprehensive Achievement Monitoring (CAM) - Reading
Richard Zweig Associates, Inc.
Fountain Valley Teacher Support System - Mathematics
Fountain Valley Teacher Support System - Reading
Scholastic Testing Service
Analysis of Skills - Mathematics
Analysis of Skills - Reading
Science Research Associates
Mastery: An Evaluation Tool - Mathematics
Mastery: An Evaluation Tool - SOBAR

as the proportion of possible points earned by a CRT and separate scores for the two weighting systems were reported for each CRT. The rating scores obtained for each CRT are reported in Table 2. In order to maintain publisher anonymity, the ratings are listed in random order and do not correspond with the alphabetical listing of publishers found in Table 1.

see page 17 for Table 2

As can be seen from Table 2, there is considerable variation in the scores each CRT earned using the two different weighting scales. When rated within the classroom resource context, the highest percentage obtained was 80 percent, with a low of 43 percent. Within the program evaluation context, the spread in the range of high and low percentages was quite similar, being 83 percent and 43 percent, respectively. A tentative conclusion, based on a review of the percentages earned by a CRT in these two contexts, might be that in trying to achieve both the classroom resource and program evaluation functions, CRTs are, for the most part, only marginally fulfilling each purpose.

Conclusions

In this study fourteen criteria used to evaluate CRTs were assigned two sets of weights reflecting the characteristics of CRTs designed as a classroom resource or for program evaluation. Currently available CRTs were rated against the criteria using each set of weights. A comparison of the scores obtained with each weighting system yielded significant differences. The findings of this study support the view that the same

criteria cannot be used for all purposes and that therefore, a CRT must be developed, validated, and evaluated in terms of the purpose for which it is intended.

Table 2

SCORE (expressed in percents) CLASSROOM RESOURCE PROGRAM EVALUATION	
65 %	43 %
65	43
74	66
51	74
80	80
77	83
57	54
66	54
71	80
63	80
57	54
57	51
77	71
60	63
51	77
54	77
49	51
51	49
46	51
49	49
43	51
43	51
54	74
54	74
57	69
49	69
49	66
43	57

References

Alkin, M. (Ed.) et al. CSE Monograph No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Anatasi, A. Psychological testing (3rd ed.). New York: Macmillan and Co., 1968.

Baker, R. Measurement considerations in instruction product development. Paper presented at the Conference on Problems in Objectives-Based Measurement, Center for the Study of Evaluation, University of California, 1972.

Buros, O. (Ed.). The mental measurements yearbook. Highland Park, New Jersey: Bryphon Press, 1972.

Glaser, R. & Nitko, A. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Hoepfner, R. et al. CSE elementary school test evaluations. Los Angeles: Center for the Study of Evaluation, University of California, 1970.

Klein, S. & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. ERIC/TM Report 26. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement and Evaluation, 1973.

Kosecoff, J., Klein, S. & Morgan, P. The useability, marketability, and technical excellence of some currently available criterion-referenced test systems. In preparation, 1976.

Kosecoff, J. & Fink, A. Feasibility of using criterion-referenced tests in the study of the sustaining effects of compensatory education. Santa Monica: System Development Corporation, 1976.

Kosecoff, J. & Fink, A. Evaluation Primer. Book in preparation, 1976.

Nafziger, D. et al. Tests of functional adult literacy: An evaluation of currently available instruments. Portland, Oregon: Northwest Regional Educational Laboratory, 1975.

Popham, W. IOX Catalog, 1976. Available from Instructional Objectives Exchange, Box 24095, Los Angeles, California, 90024.

Popham, W. & Husek, T. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6(1), 1-9.

Skager, R. Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assembly and interpretation. Paper presented at the annual American Educational Research Association meeting, New Orleans, 1973.

Wilson, H. A judgmental approach to criterion-referenced testing, CSE Monograph No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.